Report for 2003NJ48B: Automated Identification and Quantification of VOCs Using Electronic Nose Systems

There are no reported publications resulting from this project.

Report Follows

AUTOMATED IDENTIFICATION AND QUANTIFICATION OF VOCS USING ELECTRONIC NOSE SYSTEMS

PI: Robi Polikar, Ph.D., Rowan University, Electrical and Computer Engineering
Co PI: Kauser Jahan, Ph.D., P.E., Rowan University, Civil and Environmental Engineering

PROGRESS REPORT

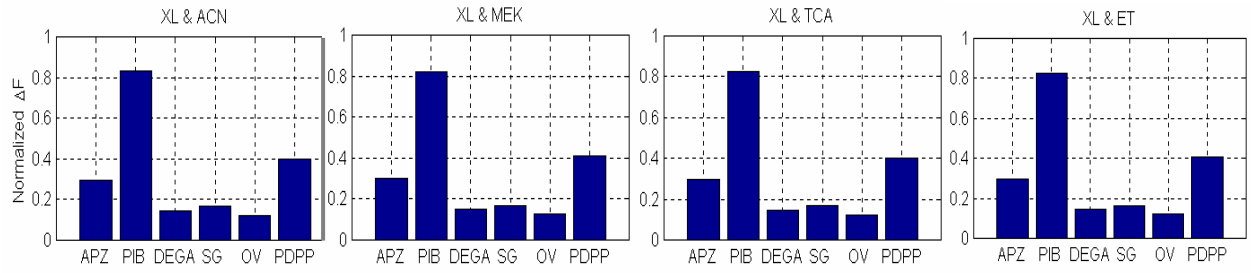
Problem and Research Objectives:

Volatile organic compounds (VOCs) are found in almost all natural and synthetic materials and are commonly used in fuels, fuel additives, solvents, perfumes, flavor additives and deodorants. Direct industrial and wastewater discharges, accidental spills of fuel products or industrial solvents, and urban runoff are the most likely sources of VOCs in surface waters. Potential health hazards and environmental degradation resulting from the widespread use of VOCs has prompted increasing concern among scientists, industry, regulatory agencies and the general public. Interest in ambient level of VOCs in the environment has increased because of the human health concerns. Of particular interest is the state of New Jersey, due to its hydrogeologically diverse, densely populated and highly industrialized nature. The Long Island – New Jersey (LINJ) study by the USGS in 1997 explored the presence of VOCs in surface and ground waters. Numerous VOCs are frequently detected in these waters.

A common method for analyzing water quality is the analytical detection of such VOCs in a laboratory setting. This method is quite sensitive, accurate and repeatable; however it requires advanced, expensive and bulky equipment, such as gas chromatographs or mass spectrometers. In addition, since these equipment are not portable, they cannot be deployed in the field, and hence require that the samples be transferred to a laboratory. Measurements can also be done by human odor assessors for certain odorous contaminants; however, such measurements are lengthy, labor intensive, expensive and subject to large variability among panelists. Furthermore, neither of the techniques allows continuous long term monitoring, since samples must be obtained from the site and transferred to an off-site where the analysis is made.

Therefore, a chemical analysis scheme that is objective, fast, accurate, cost effective, quantifiable, and field deployable would be of invaluable benefit in assessing water quality. A recent technology that has spurred interest in measurement and detection of VOCs is the *Electronic Nose (Enose)* technology. Recently, the application of Enose systems for detecting odorous compounds in wastewater treatment plants, agricultural and landfill sites has also gained prominence. Such systems usually have an array of sensors that detect odorous compounds without reference to its chemical composition. The patterns of responses obtained by these sensors are then analyzed through an automated pattern recognition system, such as a neural network. However, most studies to date have concentrated on identification of a specific VOC. In most practical cases, the VOCs appear in a mixture with other gases, typically other VOCs. Existence of several VOCs in a mixture makes the identification task considerably more challenging, primarily due to two reasons: i) the sensors themselves usually are not very selective (which is the reason for using an array of sensors); and more importantly ii) the sensors tend to be more sensitive to one of the mixture components (dominant component) then they are to others (secondary components). The responses to secondary components are then usually masked by the responses

to dominant components, which make the pattern recognition a very difficult task. This problem is illustrated in Figure 1, which shows the responses of a six-sensor array to four mixtures of Xylene (XL), with acetonitrile (ACN), methyl-ethly ketone (MEK), tricholoroethane (TCA) and ethanol (ET). We note that the responses of the sensors, which are polymer coated quartz crystal microbalances (QCMs), are remarkably similar for all four mixtures, and hence the difficulty in identifying the individual components.



The specific goal of this project is therefore to develop an artificial neural network based automated system for objective, fast, and accurate identification of VOCs that appear in mixtures. In this preliminary work, we restrict our attention to binary mixtures of VOCs, whose measurements are made by QCM type chemical sensors.

The data available to investigators for this study were acquired previously and include the 24 binary mixtures shown below. The VOC indicated at the topof each column indicate the dominant VOC. The twelve VOCs are Acetonitrile (ACN), Acetone (AC), methyl-ethyl-ketone (MEK), Octane (OC), Hexane (HX), Ethanol (ET), Methanol (ME), Xylene (XL), Toluene (TL), 1,1,1-Trichloroethane (TCA), Trichloroethylene (TCE), and 1,2-Dichloroethane (DCA).

Octane	**Xylene**	**Toluene**	**TCE**	**Ethanol**
OC & ACN	XL & ACN	TL & ACN	TCE & TCA	ET & ACN
OC & ET	XL & ET	TL & ET	TCE & MEK	ET & MEK
OC & MEK	XL & MEK	TL & MEK	TCE & TL	ET & HX
OC & TL	XL & HX	TL & HX	TCE & ET	ET & TCA
OC & TCA	XL & TCA	TL & TCA	TCE & HX	

Sensors were exposed to these mixtures at all combinations of 150, 300, 500 and 700 parts per million (ppm), giving 16 combinations of concentrations for each of the 24 mixtures listed above (that is, 150 and 150, 150 and 300, 150 and 500, 150 and 700, 300 and 150,..., 700 and 500, 700 and 700 ppm). The concentration information will first be removed from all patterns by normalizing with respect to amplitude, so that identification can be made objectively based on pattern. Once the identification is obtained, the concentrations can then be determined from calibration curves, since sensor responses are linear with respect to concentration.

Methodology:

The sensor used for determining the reactions of the VOCs in the air is an array of six quartz crystal microbalances. Each of the six microbalances is coated with a unique polymer film that reacts with the VOCs. When the coated crystal comes in contact with the VOC molecules, the molecules are deposited on the crystal surface, which then causes a measurable change in the resonant frequency of the crystal. The coatings are selected to maximize this frequencychange

for the target VOCs. The coating on the sensors used in this study were: Apiezon (APZ), Poly(isobutylene) (PIB), Poly(diethyleneglycoladipate) (DEGA), Sol-gel (SG), Poly[bis(cyanoallyl)polysiloxane)] (OV), and Poly(diphenoxylphosphorazene) (PDPP). Each of the unique polymers will react differently with the VOC mixtures and hopefully provide discriminating information for each of the different mixtures.

The automated classification system is designed as a two stage approach that attempts to classify the dominant VOC in the mixture and then uses that information to classify the secondary component. In order to facilitate the classifier's operation, ***separability algorithms*** are being considered as a preprocessing stage to accentuate the minor differences among response patterns of different VOCs. We are considering several existing algorithms, as well as developing alternatives that may work well for this particular application. Those to be explored are defined below.

- **Principal Component Analysis (PCA)**
  - This well-established algorithm attempts to find the values that project the data onto new axes where the variances of the data are greatest. Typically used to reduce the dimensionality of the problem [1].
- **Fisher Linear Discriminant (FLD)**
  - This algorithm, also well established, tries to find the projection that maximizes the discrimination between different classes of the data in a lower dimensional space [1].
  - The projection will minimize the intracluster distance (a measure of similarity of the response patterns corresponding to the same VOCs), while at the same time maximizing the intercluster distance (a measure of similarity of the response patterns corresponding to different VOCs).
- **Feature Range Stretching (FRS)**
  - Currently being developed, this algorithm adjusts the numerical ranges of pattern responses: when data values for a feature all fall in a narrow range, this algorithm maps those values to a range of [0 1] to help spread that data and make identifying classes easier [2].
- **Nonlinear Cluster Transform (NCT)**
  - Also currently being developed, this algorithm attempts to physically separate patterns from each other, without changing the dimensionality of the problem. The NCT algorithm finds a vector for each class along which all patterns are projected so that patterns of different classes are well separated from each other [2].

After application of one or more of these algorithms to the data, an automated classifier is required to determine the dominant and secondary components of the VOCs. As mentioned above, this will be done in two stages: first a neural network will be trained to identify the dominant VOC in the mixture. After this classification the data instance (the response pattern) will be passed along to one of five specialized classifiers trained specifically for the classification of the secondary VOC, given one of the five dominant VOCs. These specialized classifiers are trained on a subset of the data that contains only instances from those mixtures with a unique dominant VOC.

A series of classifiers are being explored to find a classifier that will work best for each of the stages of the classification. While one type of classifier may work for the dominant classification, a different type may be required for the secondary, so those options will be explored.

Each of the classifiers under consideration is described below along with some of the advantages or disadvantages of that approach.

- **Multi Layer Perceptron (MLP) Neural Networks**
  - By far the most commonly used and popular neural network architecture. It consists of an input layer and an output layer with one or more hidden layers in between, where each layer itself consists of a series of information processing elements, called nodes. Each node in a layer is connected to every node in the next layer through "weights", and it is these where the knowledge resides [3].
  - Each of the nodes has a nonlinear activation function associated with it, and the output value of the node is based upon the output of that function in relation to its input values. The nonlinear activation function allows the classifier to find decision boundaries between classes that are not linearly separable.
  - If the hidden layers do not have the proper number of nodes, the classifier will not be able to learn the boundary between the classes or it will over fit and not be able to classify data it has not seen accurately.
- **Radial Basis Function (RBF) Neural Network**
  - Similar to the architecture of a Multi Layer Perceptron, but the activation functions is a radial basis functions and there is only one hidden layer [3].
  - This network attempts to map the inputs from a nonlinearly separable feature space to one that is linearly separable.
  - This network is more suited for function approximation, but it has been proven to be a universal approximator, so it can be used as a classifier.
- **General Regression Neural Network (GRNN)**
  - A special case of the Radial Basis Function Neural Network, with the only difference being how the weights on the output layer of the network are determined [2].
  - Uses a statistical function approximation scheme known as nonlinear regression analysis.
  - This network does not require iterative training, so training the network becomes a less time consuming task.
- **Probabilistic Neural Network (PNN)**
  - A neural network with only one hidden layer. Each node in the hidden layer corresponds to one training data instance [1].
  - This network learns only those instances it has seen and classifies based upon how a new instance relates to those it has seen
  - For large training data sets, this algorithm requires large amounts of memory for storage of the network.
- **Learn ++**
  - An in-house developed meta-classifier that combines multiple classifiers through weighted majority voting. Learn++ seeks an improved prediction accuracy compared any single-classifier system.
  - Based upon the principle that a series of weak classifiers appropriately combined into an ensemble can perform better than one strong classifier [4].
  - Classifiers are trained using a strategic procedure where consecutive classifiers are trained to focus on those patterns that were misclassified by the previously trained classifiers.

- Each classifier is given a weight based upon how it performs on the training data. Once the system is fully trained, a response pattern is presented to all classifiers and a weighted vote is taken based on the output of each classifier and its weights. The class with the greatest weighted vote is the output of the classifier.

Work is already in progress and in all of these fronts. The specific details of the classifiers will be provided in subsequent reports as classification performance results become available.

References

[1] Duda, R., Hart, P., and Stork, D. Pattern Classification, Second Edition. John Wiley & Sons, 2001

[2] R. Polikar, Algorithms for enhancing pattern separability, feature selection, and incremental learning with applications to gas sensing electronic nose applications, *Ph.D. Dissertation,* Iowa State University, Ames IA , 2000.

[3] D.R. Hush and B.G., Horne, "Progress in supervised neural networks," *IEEE Signal Processing Magazine*, vol. 10, no.1, pp.8-39, 1993.

[4] R. Polikar, L. Udpa, S. Udpa, and V. Honavar, Learn++: An incremental learning algorithm for supervised neural networks, *IEEE Transactions on System, Man and Cybernetics (C), Special Issue on Knowledge Management*, vol. 31, no. 4, pp. 497-508, 2001.